Machine Learning Techniques for Detecting Identifying Linguistic Patterns in News Media

A Samuel Pottinger[1]

[1]Data Driven Empathy LLC, USA
{*sam@datadrivenempathy.com*}

June 4, 2019

Abstract

An article's tone and framing not only influence an audience's perception of a story but may also reveal attributes of author identity and bias. Building upon prior media, psychological, and machine learning research, this neural network-based system detects those writing characteristics in ten news agencies' reporting, discovering patterns that, intentional or not, may reveal an agency's topical perspectives or common contextualization patterns. Specifically, learning linguistic markers of different organizations through a newly released open database, this probabilistic classifier predicts an article's publishing agency with 74% hidden test set accuracy given only a short snippet of text. The resulting model demonstrates how unintentional 'filter bubbles' can emerge in machine learning systems and, by comparing agencies' patterns and highlighting outlets' prototypical articles through a new open source exemplar search engine, this paper offers new insight into news media bias.

1 Introduction

An author's language patterns influence how readers will perceive them and their message. For example, psychological and linguistic studies demonstrate how subconscious 'asymmetries' in word selection can reveal attributes of a speaker's identity and viewpoints ("The Linguistic Intergroup Bias As an Implicit Indicator of Prejudice"; "Inferring Identity From Language"; *Harvard Dialect Survey*). In addition to these unintentional phenomena, speakers may also consciously introduce bias while making an argument, influencing opinion through deliberately selected modifiers or framing (*Framing: How Politicians*

Debate). For instance, a defender of a welfare program may call its supporting taxes 'a small necessary burden' while their opponent might attack the same as 'wasteful and stifling.' This study's model examines these characteristic linguistic markers to detect patterns separating each agency from the rest, highlighting what those markings say both about news organizations and the machine learning-rich platforms disseminating their work.

1.1 Prior work

Related work in media studies, machine learning, and psychology inform this paper's modeling. To begin, prior research scores news publishers' audience ideological leaning and examine how trust in agencies differs demographically and politically (*Political Polarization & Media Habits*; *American views: Trust, media and democracy*). Bridging audience to content, earlier research also finds consumption bias towards material that agrees with a reader's pre-existing opinions ("Looking the Other Way"). In addition to investigation of audience behavior, previously published text mining techniques uncover coverage bias towards political entities within the context of specific events and infer political orientation from social media posts ("Media coverage in times of political crisis: A text mining approach"; "Predicting political affiliation of posts on Facebook"). Meanwhile, prior Bayesian analysis and more recent neural network based approaches can determine authorship of political or scientific writing (*The federalist: inference and disputed authorship*; "Authorship Attribution with Convolutional Neural Networks and POS-Eliding"). Finally, as this study earlier references, psychological human studies also show that perception of subtle patterns (like linguistic inter-group biases) allow audiences to infer attributes of the speaker's identity and beliefs ("Inferring Identity From Language"; "The Linguistic Intergroup Bias As an Implicit Indicator of Prejudice"). This study extends this prior work by learning linguistic patterns that identify a news agency across a large body of work despite variations in authorship and subject matter.

1.2 Hypotheses

This study uses the following three hypotheses to investigate news media linguistic patterns:

- Given article titles, this paper's method can predict publishing agency better than chance.
- In comparison to using titles which are often very short, this study's model can better predict the originating agency using article descriptions.
- Using an example set of topics, this method's performance increases when filtering for articles discussing a common subject by isolating different treatment of similar issues.

doi:10.20944/preprints201906.0051.v1

2 Methods and Materials

A probabilistic predictor for publishing agency investigates journalistic patterns by learning from approximately 49,000 articles to generate scores from 0 to 1 for each of the ten agencies examined. In this neural network-based classifier, 0 represents low likelihood that an article came from a particular agency and 1 represents high likelihood (*The Softmax Function, Neural Net Outputs as Probabilities, and Ensemble Classifiers*). The agency of highest score for an article becomes the model's overall prediction.

2.1 Data

A program run daily uses BeautifulSoup and Requests to capture public Really Simple Syndication ('RSS') feeds from the BBC News ('BBC'), Breitbart, CNN, Daily Mail, Drudge Report ('Drudge'), Fox News ('Fox'), New York Times ('NYT'), NPR, Vox, and Wall Street Journal ('WSJ') (*Beautiful Soup*; *HTTP for Humans*). RSS feeds are built for machine parsing and, among many uses, inform search engines of new content availability (*Are There Any SEO Benefits With Having an RSS Feed?*). This ideologically broad sample of news sources are persisted into a SQLite database with Pandas and Numpy providing data manipulation (*SQLite*; *Python Data Analysis Library*; *NumPy*). The code for gathering these data are released under a permissive open source license (*datadrivenempathy/who-wrote-this-news-crawler*). The resulting 48,997 unique articles from January 29, 2019 to May 27, 2019 are divided into training, validation, and test sets of 80%, 10%, and 10% respectively. This study publishes its dataset under a creative commons license (*Machine Learning Techniques for Detecting Identifying Linguistic Patterns in the News Media*; *Who Wrote This Data*). Note that article title and description come from the news agency itself with the later often containing article snippets or matching preview text on the source's website. Additional information is available at the dataset's webpage (*Who Wrote This Data*).

2.1.1 Abuse and standards

These data are collected in compliance with the Robots Exclusion Protocol by observing robots.txt files, industry-standard instructions which convey to systems like this crawler what data is or is not permissible to automatically process (*A Method for Web Robots Control*). This helps prevent this study from abusing services while building an index for its demonstrational search engine.

2.2 Neural network architecture

This study compares feed-forward (FF) and Long-Short Term Memory (LSTM) neural networks ("Long Short-Term Memory"). FF uses an input boolean vector encoding if a word is found within a document or not as described in Formula 1. These one dimensional vectors are processed through two fully connected 'dense' layers of variable length (using ReLUs) and a final softmax layer for output (*Deep Learning using Rectified Linear Units (ReLU)*; *The Softmax Function, Neural Net Outputs as Probabilities, and Ensemble Classifiers*). Meanwhile, the LSTM architecture features an integer encoding of the input words as a sequence where each unique word from the corpus has a unique integer index to which it is converted. These one dimensional vectors are fed into an encoding layer converting those word indices into a one-hot form before forwarding them to a variable number of LSTM units whose output are processed through a final softmax layer.

$$vector_{document}[i_{token}] = \begin{cases} 1 & \text{if } token \in document \\ 0 & \text{otherwise} \end{cases}$$

Formula 1: Occurrence Vector Formulation

Though both offer advantages, LSTMs often better fit text-based tasks as the recurrent structure can take advantage of the 'ordered' nature of language whereas FF networks treat text as unordered tokens (*Sequence Classification with LSTM Recurrent Neural Networks in Python with Keras*). However, prior work does see FF outperform LSTM in some circumstances (*When Recurrent Models Don't Need to be Recurrent*). To contrast these structures, Figure 1 summarizes both architectures.

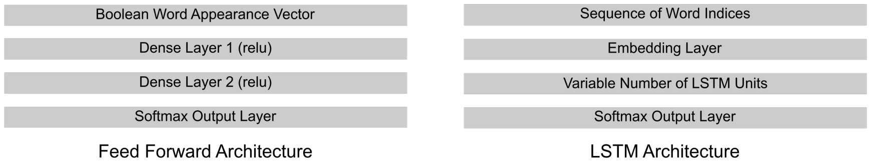


Figure 1: Neural network template architectures

2.3 Regularization

To prevent overfitting where performance in training exceeds that in a hidden validation set, regularization techniques in both FF and LSTM improve predictive generalization. First, this paper varies L2 activation regularization within interior layers of both network types (dense layers and LSTM units) before also varying dropout rates within the same layers ("Feature Selection, L1 vs. L2 Regularization, and Rotational Invariance"; "Dropout: A Simple Way to Prevent Neural Networks from Overfitting"). Additionally, as uncommon words

doi:10.20944/preprints201906.0051.v1

may uniquely identify articles, the number of words included in the encoding provide a form of regularization. Therefore, the frequency of a word is calculated across the entire dataset and only the ‘top n’ words are included in the encoding (with multiple values of n explored).

2.4 Other preprocessing

Some articles ‘leak’ information about the publisher, preventing actual extraction of linguistic patterns. This includes, for example, hyperlinks to the agency’s website or a publisher referring to itself (an article mentioning it is part of a Vox series reveals its publisher to be Vox). With this in mind, HTML elements are converted to the text visible to the user and the name of the publisher is removed from the title and description text before training. Some agency-specific features that are not part of the ‘body’ of the title or description such as agency-specific advertising text are also hidden from the neural network.

2.5 Voice similarity

Incorrectly classified articles may indicate that two news agencies exhibit similar ways of speaking, helping explain model behavior and the nature of media linguistic patterns. For example, a large number of articles from the BBC classified as published by NPR may indicate that the two share similar characteristics. With that motivation in mind, this study offers bidirectional similarity calculations through a Jaccard Index-like measure (“The Distribution Of The Flora In The Alpine Zone.1”). In Formula 2, $s1$ refers to the set of articles from source 1, $s2$ refers to the set of articles from source 2, $classified(s1, s2)$ refers to the set of articles incorrectly classified as being from source 2 while actually being from source 1, and $similarity(s1, s2)$ refers to the bidirectional similarity between $s1$ and $s2$. A larger $similarity(s1, s2)$ could indicate a greater overlap in patterns.

$$similarity(s1, s2) = \frac{|classified(s1, s2)| + |classified(s2, s1)|}{|s1| + |s2|}$$

Formula 2: Bidirectional Similarity

2.6 Comparison to prior work

In order to discuss findings within the context of prior work, this paper presents additional formulations for comparing results to other pre-existing measures.

2.6.1 Formulation for comparison with Pew

Pew audience ideology scores provide a conservative to liberal scale for news agencies (*Political Polarization & Media Habits*; "Looking the Other Way"). However, Formula 2's method presents some issues when comparing Pew's work to this model. First, Formula 2 does not provide a one dimensional scale, requiring this study to measure distances from a reference agency in this paper's model to all other outlets to create a single axis of comparison. Breitbart in particular provides a convenient reference as it is the most conservative source this paper examines according to Pew (*Political Polarization & Media Habits*). Second, values generated using Formula 2 will be close to 0 if $s1! = s2$ and close to 1 if $s1 = s2$. This causes a 'crowding' of values on the extremes of a distance measure when agencies' distance to themselves are included. That in mind, Formula 3's alternative distance from source 1 ($s1$) to source 2 ($s2$) calculates the percent of incorrectly classified articles from $s1$ that were predicted as $s2$. In contrast to Formula 2, this affords comparison to Pew scores but is only suited to 'directional' tasks (no guarantee that $distance(s1, s2) = distance(s2, s1)$).

$$distance(s1, s2) = \begin{cases} 0 & \text{if } s1 = s2 \\ 1 - \frac{|classified(s1,s2)|}{|s1|-|classified(s1,s1)|} & \text{otherwise} \end{cases}$$

Formula 3: Directional Distance Function

2.6.2 Formulation for comparison with sentiment polarity

To understand model behavior, this study also contrasts results to a pre-trained sentence polarity model, examining model relationship to not only agencies' average title polarity score but also the percent of articles from an agency whose titles exhibit 'substantial overall' positive or negative polarity (found by filtering for titles with polarity averages < -0.3 or > 0.3) (*Simplified Text Processing*).

2.7 Understanding feature importance

To understand the types of patterns that this model identifies, this study presents findings through Local Interpretable Model-agnostic Explanations (LIME) ("'Why Should I Trust You?': Explaining the Predictions of Any Classifier"). This method can quantify the importance of features (words) to model predictions.

2.8 Finding exemplars and examining topics

Qualitatively exploring the nature of news media patterns, the final softmax layers' 0 to 1 scores indicate likelihood that an article came from a particular news agency such that filtering to the top 'n' documents sorted by an agency's score in descending order identifies news articles that are most like their publishing source. This reveals, for example, which NPR articles are the most prototypical of NPR coverage within the model. Similarly, one can find cross-agency exemplars and sorting Fox articles by their Breitbart score allows one to find the most 'Breitbart-like' articles published by Fox, revealing possible similarities in voice. Finally, filtering for exemplars within a topic affords comparison of news sources within a single issue or event, subject matter exemplars this paper explores through an example topic set that the author believes a priori may provoke interesting comparisons: school, climate, democrat, Trump, and immigration. Though this paper uses this set below, an open source application enables investigation of other topics (*Who Wrote This App*).

2.9 Supplemental resources and reproducibility

Multiple configurations for neural networks are considered through a newly released open source system and, in addition to code and data in Code Ocean for reproducibility, this study supplements this methods section via Protocols.io (*datadrivenempathy/who-wrote-this-training*; *Machine Learning Techniques for Detecting Identifying Linguistic Patterns in the News Media*; "Machine Learning Techniques for Detecting Identifying Linguistic Patterns in the News Media").

3 Results

The study finds evidence to support all three originally posed hypotheses (restated and referred to as Hypotheses 1 through 3):

1. Given article titles, this paper's method can predict publishing agency better than chance.

2. This study's model can better predict the originating agency using article description compared to using title.

3. This method's performance will increase when filtering for articles discussing a common topic from within an example subjects set.

3.1 Model performance

Both FF and LSTM outperform the naive ('by chance') method and use of description outperforms title (Figure 2), supporting both hypotheses 1 and 2. Note that the 'naive' approach is to guess the source with the largest number of articles published (Daily Mail), resulting in a 32% accuracy. This paper later provides per agency scores and class imbalance is further discussed below.

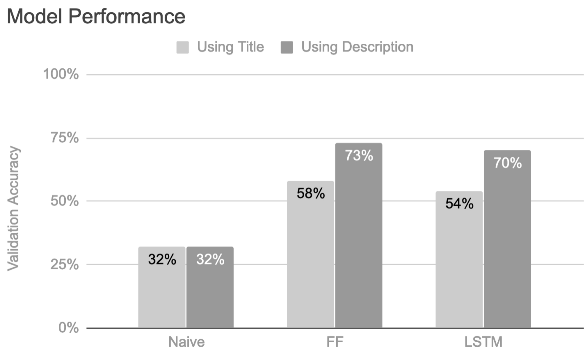


Figure 2: Model outperformed ‘naive’ approach across inputs

3.1.1 Feed-forward or LSTM

To decide between FF and LSTM, both are trained on the same corpus using a set of similar parameters (descriptions, 10,000 words, 0.01 L2 regularization, no dropout). LSTM achieves 70% validation set accuracy while FF achieves 73%, justifying a preference for FF. As cataloged, training within other parameter sets similarly sees FF outperform LSTM in this study (*README*).

3.1.2 Architectural variation

In FF, all dense layer sizes yield comparable validation accuracy as shown in Table 1. Similarly, the number of LSTM units does not drastically impact validation accuracy (16 units yield 68%, 32 yield 70%). Based on this, FF with the (32, 16) layer configuration is used to classify news agency source in subsequent evaluations as it achieves the highest validation set accuracy.

FF 1st Dense Layer	FF 2nd Dense Layer	Validation Accuracy
8	8	71%
16	8	71%
32	16	73%
64	32	73%

Table 1: Feed forward dense layer sizes and validation set performance

3.1.3 Regularization

Regularization dramatically impacts performance and moderate L2 regularization yields preferred accuracies as shown in Figure 3.
This paper also explores various dropout rates in Figure 4, observing peak validation accuracy around 0.6. That said, this study prefers the more balanced

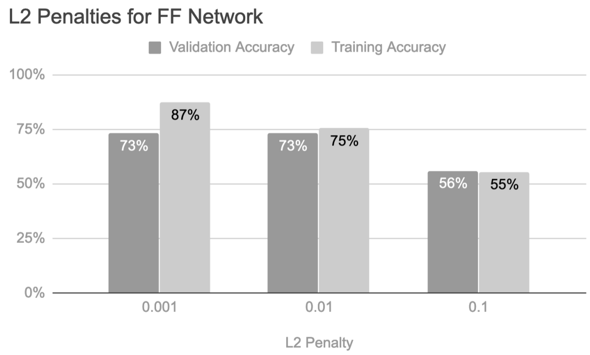


Figure 3: L2 penalties

performance of L2 regularization and uses L2 of 0.01 in subsequent comparisons.

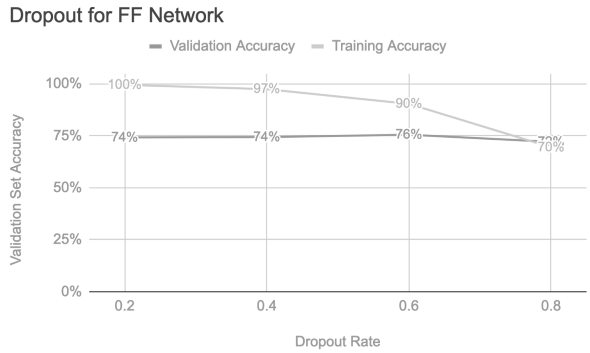


Figure 4: Dropout rates

3.1.4 Number of words encoded

The study hypothesizes that the number of words included in the encoding could provide a form of regularization as highly uncommon words may uniquely identify articles or a small set of articles. The effect of encoder size is tested as reported in Figure 5. In general, the model saw performance degradation in smaller dictionary sizes (fewer unique words encoded) and, while showing minor over-fitting, larger dictionary sizes lead to higher accuracy on the test set. However, the marginal utility of additional words appears to decrease after around 10,000 words. Specifically, as shown in Figure 5, 10,000 words provide slightly improved validation set accuracy compared to 30,000 words.

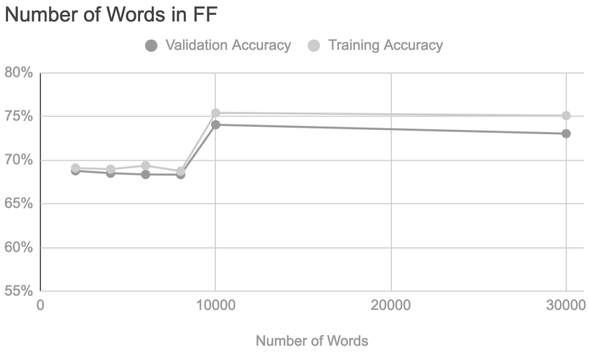


Figure 5: Performance across different number of words encoded

3.2 Selected configuration

Given empirical results, the study recommends the following:

- 32 node first dense layer, 16 node second dense layer FF.
- 0.01 L2 kernel regularization but no dropout.
- Using 10,000 top words.

The model achieves achieves a 74% test set accuracy after selecting this configuration, a result in line with its 73% validation set accuracy. This paper uses these settings in presenting results below.

3.3 Topical performance

Next, supporting Hypothesis 3, accuracy increases when comparing different organizations coverage of the same subjects (Figure 6). However, uneven performance distribution across topics may indicate that some subjects showcase differences in voice better than others. Note that this test uses accuracy across all sets (test, validation, and training) as topic filtering causes a sharp reduction in n size, especially when considering either just validation or test sets. Also, due to filtering on topic yielding low sample sizes, observe that the model is not retrained when evaluating each topic. As explored later, this may suggest that the model compares signals within the context of a topic (different agencies' 'treatments' of an issue) but the same model can perform across different subject matter.

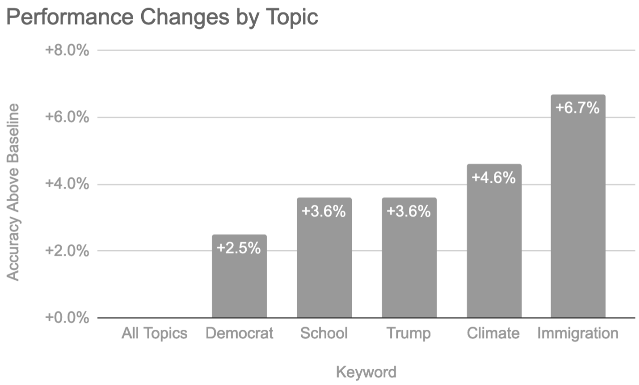


Figure 6: Performance after filtering for keywords within the example topics set

3.4 Voice similarity

Helping to understand these linguistic relationships, consider that two news agencies may have similar ways of speaking if articles from one source are often confused as having come from another source. Figure 7 shows the results of applying Formula 2 and is consistent with prior work on political ideology, demonstrating clear links between NPR, CNN, and the New York Times as well as between Breitbart and Fox (*Political Polarization & Media Habits*). That said, some edges may indicate pattern similarities less emergent from politics. Though prior work would suggest that their political leaning differs, Fox's strong connection to CNN indicates that the 'voice' in which they deliver their different message is still itself similar (*Political Polarization & Media Habits*). This is explored further below.

3.5 Quantitative understanding of patterns

To understand the types of patterns identified, this study compares this model's behavior to related work and calculates feature (word) importance.

3.5.1 Ideology and polarity

Using Formula 3, Figure 8 describes a modest correlation between the Pew score and the distance from Breitbart in this model ($R^2 = 0.3$). This relationship suggests that, while this classifier may interpret political signals, model behavior is poorly explained by political lean alone.

Similarly, Figure 9 depicts a modest correlation between model Breitbart distance and percent of titles of substantial negative polarity ($R^2 = 0.4$). That said, positive polarity percent ($R^2 < 0.1$) and average polarity score ($R^2 = 0.2$)

doi:10.20944/preprints201906.0051.v1

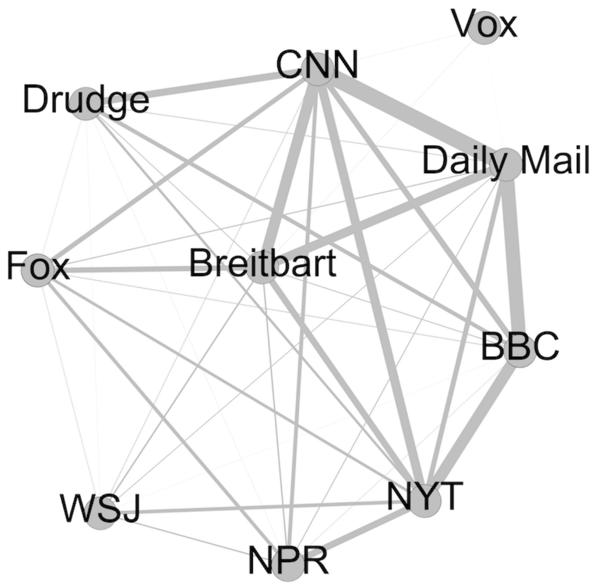


Figure 7: News agency graph with line width proportional to similarity score

do not correlate as well with Breitbart model distance. Therefore, while polarity and politics likely play a role in the patterns identified by this model, these findings suggest that neither can explain the model's overall behavior.

3.5.2 Word importance

As polarity and ideology fail to fully explain model behavior, this study next evaluates the importance of different words using LIME ("'Why Should I Trust You?': Explaining the Predictions of Any Classifier"). However, using the prototypical articles' described in Section 2.8, even the ten most 'important' words identified per article only weakly influence predictions (Figure 10).
Indeed, 98% of words examined shift an agency's score by less than 0.2 (57% under 0.05) despite this prototypical set's agency scores nearing 1. Furthermore, as published in supplemental material, LIME observes the model using logical candidates for discrimination (Donald, growth, wounded, police, federal, 'England' rugby team, etc.) even if no token achieves overwhelming power (*Who Wrote This Template Workbook*). This highly dispersed token importance dis-

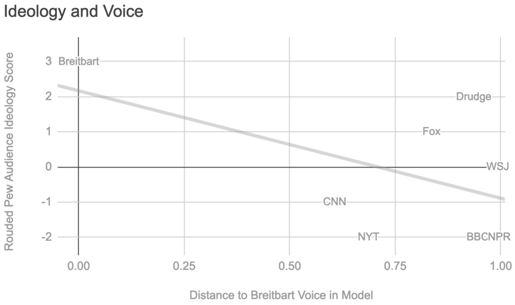


Figure 8: Pew audience ideology score and Breitbart distance

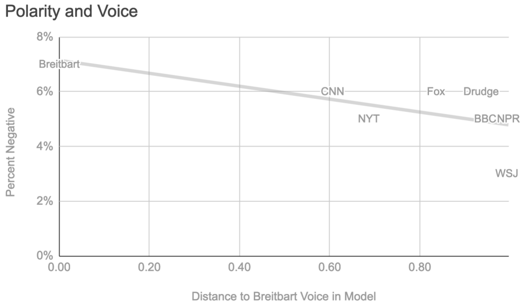


Figure 9: Percent articles with negative polarity and Breitbart distance

tribution implies that this model behavior is complex beyond simple word or topic probabilities, possibly indicating that the model is interpreting something like contextual sentiment towards topics or framing devices. This study further explores this in the qualitative results below.

3.6 Qualitative understanding of patterns

To give context to quantitative exploration, prototypical articles provide qualitative understanding of these patterns as they manifest in individual stories.

3.6.1 Agency exemplars

Starting with full corpus exemplars, consider Table 2's 'prototypical' articles, stories most 'like' their publishing agency within this paper's model. These representative stories help yield observations like how Breitbart discusses figures from other outlets and how, echoing earlier correlational findings, it uses highly polarized language. Of course, mirroring token importance above, Table 2 may

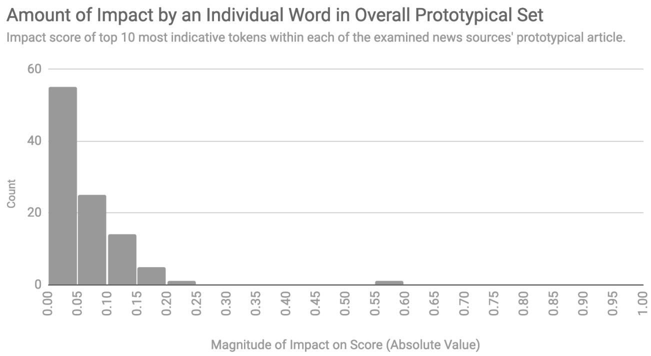


Figure 10: LIME analysis of prototypical articles

also reveal some possible topical bias like in BBC's reference to England's rugby team. That said, in addition subject matter discrimination, this collection also highlights characteristic vantage points for some agencies (Fox discussing violent crime and the Wall Street Journal's economic focus) that emerge again below. Still, consistent with the quantitative results, none of these features (sentiment, topical bias, and framing) seem to alone describe this exemplar set.

Source	**Exemplar Title**
BBC News	Jones to bring in expert in bid to fix England's mental weakness under pressure
Breitbart	Mark Levin: ’Hate-America Democrats Passed a Resolution Telling You That America Sucks‘
CNN	The move shows a potential growing threat to the President and those in his orbit from probes by the Manhattan US Attorney's office
Daily Mail	Children behind half of London knife crime as machete is sold for just £19
Drudge Report	Pelosi Warns Dems: Stay in Center; Trump May Contest Election Results...
Fox News	The Latest: Vegas police: Wounded robbery suspect has died
NPR	Retired Military Officers Urge Caution In Proposed Diplomatic Spending Cuts
New York Times	French Raise a Glass to a Health Warning About Too Much Wine
Vox	Google employees walked out for the right to sue their bosses. Now they're taking the fight to Congress.
Wall Street Journal	How Bad Is the China Slowdown? U.S. Companies Offer Some Answers

Table 2: Overall exemplar articles per agency

3.6.2 Topical exemplars

Prototypical articles can also isolate differences in coverage in the context of a particular topic, sharpening a view into the model's learned trends. For example, Table 3 displays prototypical stories discussing climate. While again no single attribute of 'voice' seems to define all of the sources, some observations from that set which may reveal framing devices:

- Breitbart frames climate change within context of political power.
- Fox's exemplar places the topic within a political frame whereas NPR places it within a scientific one.
- Wall Street Journal situates the issue within an economic context, consistent with its economic focus.
- Both BBC News and the Daily Mail discuss climate protesters but one puts an emphasis on violence.

Source	Exemplar Title
BBC	The Papers: Climate protests and Trump probe
Breitbart	Warren: Climate Change, Gun Violence, Student Loan Debt Constitute for National Emergency Declaration
CNN	John Avlon speaks the cold truth about climate change
Daily Mail	Dramatic moment police DRAG two climate change protesters along the street
Drudge Report	Climate-first...
NPR	The Role Climate Change Plays In Weather Extremes
Fox	Trump pokes fun at Klobuchar's climate-change stance as she announces candidacy in snow
New York Times	Nonfiction: Two New Books Dramatically Capture the Climate Change Crisis
Vox	Amazon says it's a leader on fighting climate change. 5,000 employees disagree.
Wall Street Journal	Glencore, the King of Coal, Bows to Investor Pressure Over Climate

Table 3: Prototypical articles discussing climate

3.6.3 Cross-agency exemplars

Finally, to qualitatively understand why some agencies exhibit connections in Figure 7, Table 4 examines articles from one agency showing patterns of another using the method from Section 2.8. The high polarity of CNN / Fox articles contrasts low polarity economic focus of the Wall Street Journal / NPR articles. Though prior work would not expect a connection between CNN and Fox or NPR and the Wall Street Journal, these cross-exemplars highlight possible non-political connections shared between agencies (*Political Polarization & Media Habits*).

Voice of Source	Actual Source	Title
CNN	Fox	AP FACT CHECK: Trump blames media for his McCain rant
Fox	CNN	Gunman in Quebec mosque shooting sentenced to life in prison
Wall Street Journal	NPR	Fed Changes Course, Holds Off On Raising Interest Rates
NPR	Wall Street Journal	U.S. Bets on China's Special Envoy

Table 4: Selection of cross-agency articles

3.7 Other observations

This study wishes to highlight final observations within results before discussion.

3.7.1 Per-agency performance

Figure 11 shows some like Vox have 'high separability' whereas some like Fox saw more confusion.

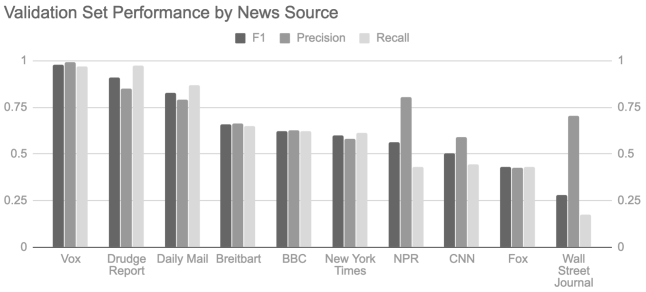


Figure 11: Validation set performance per agency

That said, a weak correlation of F1 score to sample size ($R^2 = 0.2$) suggests added data aids learning but class imbalance does not bias towards more strongly represented agencies (Figure 12). Therefore, Figure 11 likely sees feature discernibility lead some agencies to be more identifiable than others.

3.7.2 Fox News

In early training, Fox News' recall in the validation set was below 1% and, to remedy this, Fox is re-sampled (articles duplicated) so that additional instances appear in training (but not in test or validation sets). The amount of resampling (from 0% to 100%) of Fox does not impact overall validation accuracy by more than 1% but, using 100% duplication (each Fox article in the training set duplicated exactly once), Fox's recall reaches 43%. This study does not resample other agencies as performance is not as disparate for any other outlet. Depending on use case, derived work may consider this resampling optional.

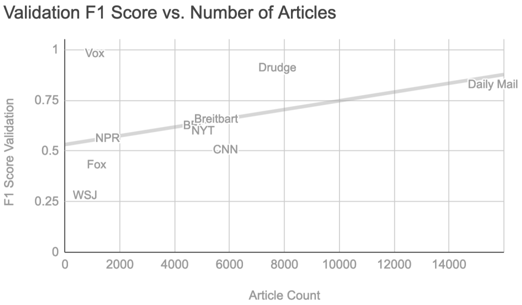
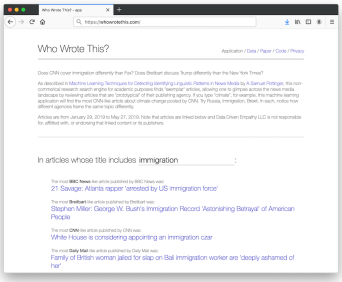


Figure 12: Validation set F1 score versus number of articles for an agency

3.8 Search engine

To enable researchers to further explore the exemplars generated by this model, a public demo search engine is made available at whowrotethis.com using the Flask framework and an 'inverted index' (*Flask (A Python Microframework)*; *Exploring Solr Internals : The Lucene Inverted Index*). The application (Figure 13) is available as open source (*datadrivenempathy/who-wrote-this-server*).


Who Wrote This? - app
https://whowrotethis.com/
Application / Data / Paper / Code / Privacy

Who Wrote This?

Does CNN cover immigration differently than Fox? Does Breitbart discuss Trump differently than the New York Times?

As described in Machine Learning Techniques for Detecting Identifying Linguistic Patterns in News Media by A Samuel Pottinger, this non-commerical research search engine for academic purposes finds "exemplar" articles, allowing one to glimpse across the news media landscape by reviewing articles that are "prototypical" of their publishing agency. If you type "climate", for example, this machine learning application will find the most CNN-like article about climate change posted by CNN. Try Russia, Immigration, Brexit. In each, notice how different agencies frame the same topic differently.

Articles are from January 29, 2019 to May 27, 2019. Note that articles are linked below and Data Driven Empathy LLC is not responsible for, affiltted with, or endorsing that linked content or its publishers.

In articles whose title includes immigration :

The most **BBC News**-like article published by BBC News was:
21 Savage: Atlanta rapper 'arrested by US immigration force'

The most **Breitbart**-like article published by Breitbart was:
Stephen Miller: George W. Bush's Immigration Record 'Astonishing Betrayal' of American People

The most **CNN**-like article published by CNN was:
White House is considering appointing an immigration czar

The most **Daily Mail**-like article published by Daily Mail was:
Family of British woman jailed for slap on Bali immigration worker are 'deeply ashamed of her'

Figure 13: Who Wrote This running in Firefox on Mac OS X

4 Discussion

This study's model demonstrates how linguistic patterns can not only identify individual agencies but also reveal characteristic tone and framing. Furthermore, while this neural network-based classifier may include signals of sentiment polarity and political ideology, those measures fail to fully describe these predictions which are more complex than an individual word's probabilistic relationship with a publisher. After demonstrating predictive power and information retrieval capability, this study now turns to discussion of findings.

4.1 Implications for modeling

The length of documents, number of training instances, and choices made in regularization / model architecture all impact accuracy. However, this study finds complex relationships between outcome and those inputs and structures.

4.1.1 Sensitivities to amount of input data

As described, amount of input data impacts accuracy but more data may not always yield better outcome. Most notably, description out-performs title likely because the later is longer (Figure 14). That said, increasing dictionary size does improve performance but only up to a certain point of rapidly diminishing returns and, while increased sample sizes may aid learning for some agencies, more training instances do not guarantee increased performance.

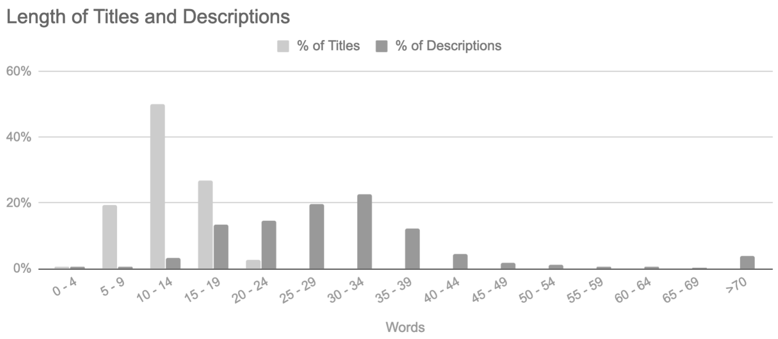


Figure 14: Lengths of inputs across all agencies

4.1.2 Decisions for regularization

This model may require regularization because articles have highly identifying word co-occurrences which enable them to be identified without extracting broader patterns. That said, weight penalties like L2 regularization outperform dropout in this model and results suggest L2 weight penalties may also improve regression in related tasks such as topic modeling or authorship assignment.

4.1.3 Feed-forward vs LSTM

FF prevails in this task possibly because, given recent research in useable context, the number of tokens per instance may simply be too short to take full advantage of LSTM's properties (*Sharp Nearby, Fuzzy Far Away: How Neural Language Models Use Context*). Note that there have been other findings of FF outperforming recurrent structures in some circumstances (*When Recurrent Models Don't Need to be Recurrent*). Beyond accuracy, FF sees notably lower wall clock training time than LSTM as cataloged (*README*).

4.2 Emergent signals of ideology

While model behavior seems broader than conservative versus liberal or sentiment polarity, some learned patterns are likely ideological in nature. For example, a learned belief around agency co-occurrence of 'economy' and 'climate' or 'Trump' and 'honorable' could be a form of 'commentary' on a publisher, an emergent 'form of speech' without explicitly ideological feature engineering. While this model seeks that linguistic bias, this could have implications for other machine learning systems desiring or requiring neutrality. Even without express instruction to extract ideology, consider that news recommenders could learn ideological-correlated features to suggest articles matching users' subjectivities, creating or reinforcing 'filter bubbles' (*Filter bubbles are a serious problem with news, says Bill Gates*). Alternatively, in 'cold start' or a non-personalized system where all users are presented the same recommendations, an unintended non-neutrality in the platform could emerge to reflect majority preferences (*The Cold Start Problem for Recommender Systems — Yuspify Blog*). Similar to prior work discussing racial or sexual disparity in algorithm performance, this unintended bias may impact discussion of technology companies' ethical responsibilities when they act as mediators of the news media ecosystem or host public discourse (*Is Facebook a publisher? In public it says no, but in court it says yes*; *Is Artificial Intelligence Racist?*). Furthermore, like earlier work in biased word embeddings, recognition of possible content-emergent bias could open avenues of remediation in both recommenders and other systems processing this kind of text ("Man is to Computer Programmer As Woman is to Homemaker? Debiasing Word Embeddings").

4.3 Implications for news media

Americans' trust in different news agencies is neither high nor uniform across demographic group or political affiliation but, while many studies, news stories, and even commercial products examine differences in coverage on the axis of politics, this study suggests that journalistic voice is more complex than ideological lean (*American views: Trust, media and democracy*; *Political Polarization & Media Habits*; "Looking the Other Way"; *How to Escape Your Political Bubble for a Clearer View*; *Balanced news via media bias ratings for an unbiased news perspective*). In addition to recommending future news media studies consider non-political angles, this study in particular calls for work in determining if non-political patterns influence news consumption behavior because, if these other attributes of 'voice' like framing or polarity influence consumption in ways similar to politics, an article's addressable audience may not only be limited by perceptions of its political bias but by audience affinity for the other linguistic markers with which it is written, further dividing the modern media landscape beyond politics alone (*American views: Trust, media and democracy*; "Looking the Other Way"). In short, if voice 'mediates' the way an article is received, these linguistic patterns not only present a challenge to a journalist attempting to reach an audience but non-political signals may further shrink the number of avenues through which 'facts' are established for a population. This is discussed further in future work.

5 Future Work

This study suggests multiple avenues for future work both within machine learning and through other methods of inquiry.

5.1 Class imbalance

Though this study does take some steps to address class imbalance when it impacts performance, future studies could explore instance weighting and resampling to confront the fact that some agencies publish more work. That said, this paper suggests that an increased number of instances may or may not improve performance and some infrequent classes do see high performance in the current model iteration.

5.2 Full article length

This study does not attempt classification using full article text which could potentially improve performance. Still, given that the number of words included in an encoding offers only a diminishing return after a certain point, this study suggests that regressing on the full article may not necessarily improve performance substantially.

5.3 Testing voice preferences

As discussed, this paper cannot empirically establish that the identified patterns increase appeal to certain audiences and, in the case of non-political patterns, it is possible but not necessarily true that these markers impact media consumption ("Looking the Other Way"). Future work could close this gap by exploring the interaction of favorability and voice in different populations similar to prior

doi:10.20944/preprints201906.0051.v1

work but outside an explicitly political lens, comparing framing devices (economic vs political) or polarity for example ("Looking the Other Way").

5.4 Topic modeling and sentiment analysis

This paper simply filters for keywords within a title when exploring a topic and brief experiments with Latent Dirichlet Allocation and Tf-Idf did not result in sensible topic modeling yet ready for publication ("Latent dirichlet allocation"; "Document clustering: TF-IDF approach"). Furthermore, this study only uses pre-existing sentiment models. Therefore, it is possible other studies could create more performant sentiment or topic modeling system with tailoring to this dataset, possibly providing alternatives to this paper's example subject set.

6 Conclusions

Learning from publicly available RSS feeds for ten news organizations, this machine learning method uncovers different agencies' linguistic characteristics through a neural network-based model capable of identifying the publisher of an article with 74% test set accuracy. This probabilistic classifier allows for quantitative contrasting of different organizations' patterns and empirical findings provide recommendations for related tasks both on this study's newly released open database and in similar news media datasets. Furthermore, an 'exemplar' search engine using the learned model enables researchers to qualitatively uncover new insight into organization's coverage. Finally, ending with commentary rooted in previous linguistic, psychological, and media studies research, discussion highlights possible implications of these patterns for the media landscape and, in demonstrating a mechanism for inadvertent 'filter bubble' emergence, describe practical ethical implications for other machine learning efforts. That said, in addition to recommending investigation outside an explicitly liberal versus conservative lens, this paper invites future work to further investigate observed phenomena both through machine learning methods like topic modeling and through psychological / linguistic studies.

Reproducibility and data

This study releases code under a permissive open source license and provides runnable containers via Code Ocean (*datadrivenempathy/who-wrote-this-training*; *datadrivenempathy/who-wrote-this-news-crawler*; *datadrivenempathy/who-wrote-this-server*; *Machine Learning Techniques for Detecting Identifying Linguistic*

Patterns in the News Media). Empirical results for attempted model configurations are also archived via Weights and Biases and the methods section is supplemented via Protocols.io (*README*; "Machine Learning Techniques for Detecting Identifying Linguistic Patterns in the News Media"). To respect the spirit of a publisher's wishes, data are made available under the Attribution-NonCommercial-ShareAlike 4.0 International Creative Commons license (*Who Wrote This Data*; *CNN RSS*). It is advised that derived work ensure compliance with governing restrictions around use of article information including possible applicability of 'fair use' and, especially if crawling, updated state of publishers' robots.txt files (*More Information on Fair Use*; *A Method for Web Robots Control*).

Acknowledgements

This study thanks Carey Phelps (Product Lead at Weights and Biases) for conversation that motivated this research as well as Shreyas Jayanna and David Joy for suggestions regarding presentation, investigation, and implications of these findings.

Competing Interests

Carey Phelps is a friend of AP and this study's machine learning pipeline optionally uses her employer's product (Weights and Biases) to log training information. Operation of all released code and use of released data does not require leveraging Weights and Biases.

doi:10.20944/preprints201906.0051.v1

References

Agarap, Abien Fred. 2018. *Deep Learning using Rectified Linear Units (ReLU)*. Available at http://arxiv.org/abs/1803.08375 [Last accessed on 3 June 2019].

AllSides. 2019. *Balanced news via media bias ratings for an unbiased news perspective*. Available at https://www.allsides.com/unbiased-balanced-news [Last accessed on 3 June 2019].

Bafna, Prafulla, Dhanya Pramod and Anagha Vaidya. 2016. “Document clustering: TF-IDF approach”. In: *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*. DOI: 10.1109/iceeot.2016.7754750.

Benedetti, Alessandro. 2018. *Exploring Solr Internals : The Lucene Inverted Index*. Available at https://sease.io/2015/07/exploring-solr-internals-lucene.html [Last accessed on 3 June 2019].

Blei, David M. et al. 2003. “Latent dirichlet allocation”. In: *Journal of Machine Learning Research* 3, p. 2003.

Bolukbasi, Tolga et al. 2016. “Man is to Computer Programmer As Woman is to Homemaker? Debiasing Word Embeddings”. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS’16. Barcelona, Spain: Curran Associates Inc., pp. 4356–4364. ISBN: 978-1-5108-3881-9. Available at http://dl.acm.org/citation.cfm?id=3157382.3157584.

Brownlee, Jason. 2019. *Sequence Classification with LSTM Recurrent Neural Networks in Python with Keras*. Available at https://machinelearningmastery.com/sequence-classification-lstm-recurrent-neural-networks-python-keras/ [Last accessed on 3 June 2019].

Chang, Che-Chia, Shu-I Chiu and Kuo-Wei Hsu. 2017. "Predicting political affiliation of posts on Facebook". In: *Proceedings of the 11th International Conference on Ubiquitous Information Management and Communication - IMCOM '17*. DOI: 10.1145/3022227.3022283.

CNN RSS. Available at http://www.cnn.com/services/rss/#terms [Last accessed on 3 June 2019].

de Bruijn, Hans and Gerdien de Vries. 2018. *Framing: How Politicians Debate*. Available at https://www.edx.org/course/framing-how-politicians-debate-delftx-frame101x-1 [Last accessed on 3 June 2019].

Delaney, Kevin J. 2017. *Filter bubbles are a serious problem with news, says Bill Gates*. Available at https://qz.com/913114/bill-gates-says-filter-bubbles-are-a-serious-problem-with-news/ [Last accessed on 3 June 2019].

Fortuny, Enric Junqué De et al. 2012. "Media coverage in times of political crisis: A text mining approach". In: *Expert Systems with Applications* 39.14, pp. 11616–11622. DOI: 10.1016/j.eswa.2012.04.013.

Gaspar, Huba. 2018. *The Cold Start Problem for Recommender Systems — Yuspify Blog*. Available at https://www.yuspify.com/blog/cold-start-problem-recommender-systems/ [Last accessed on 3 June 2019].

Hess, Amanda. 2017. *How to Escape Your Political Bubble for a Clearer View.* Available at `https://www.nytimes.com/2017/03/03/arts/the-battle-over-your-political-bubble.html` [Last accessed on 3 June 2019].

Hippel, William Von, Denise Sekaquaptewa and Patrick Vargas. 1997. “The Linguistic Intergroup Bias As an Implicit Indicator of Prejudice”. In: *Journal of Experimental Social Psychology* 33.5, pp. 490–509. DOI: `10.1006/jesp.1997.1332`.

Hitschler, Julian, Esther Van Den Berg and Ines Rehbein. 2017. “Authorship Attribution with Convolutional Neural Networks and POS-Eliding”. In: *Proceedings of the Workshop on Stylistic Variation.* DOI: `10.18653/v1/w17-4907`.

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. “Long Short-Term Memory”. In: *Neural Computation* 9.8, pp. 1735–1780. DOI: `10.1162/neco.1997.9.8.1735`.

Jaccard, Paul. 1912. “The Distribution Of The Flora In The Alpine Zone.1”. In: *New Phytologist* 11.2, pp. 37–50. DOI: `10.1111/j.1469-8137.1912.tb05611.x`.

Khandelwal, Urvashi et al. 2018. *Sharp Nearby, Fuzzy Far Away: How Neural Language Models Use Context.* Available at `https://nlp.stanford.edu/pubs/khandelwal2018lm.pdf` [Last accessed on 3 June 2019].

Knight Foundation. 2018. *American views: Trust, media and democracy.* Available at `https://knightfoundation.org/reports/american-views-trust-media-and-democracy` [Last accessed on 3 June 2019].

Knobloch-Westerwick, Silvia and Jingbo Meng. 2009. “Looking the Other Way”. In: *Communication Research* 36.3, pp. 426–448. DOI: 10.1177/0093650209333030.

Koster, M. 1996. *A Method for Web Robots Control*. Available at http://www.robotstxt.org/norobots-rfc.txt [Last accessed on 3 June 2019].

Lan, Haihan. 2017. *The Softmax Function, Neural Net Outputs as Probabilities, and Ensemble Classifiers*. Available at https://towardsdatascience.com/the-softmax-function-neural-net-outputs-as-probabilities-and-ensemble-classifiers-9bd94d75932 [Last accessed on 3 June 2019].

Levin, Sam. 2018. *Is Facebook a publisher? In public it says no, but in court it says yes*. Available at https://www.theguardian.com/technology/2018/jul/02/facebook-mark-zuckerberg-platform-publisher-lawsuit [Last accessed on 3 June 2019].

Loria, Steven. 2019. *Simplified Text Processing*. Available at https://textblob.readthedocs.io/en/dev/ [Last accessed on 3 June 2019].

Miller, John. 2018. *When Recurrent Models Don’t Need to be Recurrent*. Available at https://bair.berkeley.edu/blog/2018/08/06/recurrent/ [Last accessed on 3 June 2019].

Mitchell, Amy et al. 2014. *Political Polarization & Media Habits*. Available at https://www.journalism.org/2014/10/21/political-polarization-media-habits/ [Last accessed on 3 June 2019].

Mosteller, Frederick and David L. Wallace. 1964. *The federalist: inference and disputed authorship*. Addison-Wesley.

Ng, Andrew Y. 2004. "Feature Selection, L1 vs. L2 Regularization, and Rotational Invariance". In: *Proceedings of the Twenty-first International Conference on Machine Learning*. ICML '04. Banff, Alberta, Canada: ACM, pp. 78–. ISBN: 1-58113-838-5. DOI: 10.1145/1015330.1015435. Available at http://doi.acm.org/10.1145/1015330.1015435.

Numpy. 2019. *NumPy*. Available at https://www.numpy.org/ [Last accessed on 3 June 2019].

Pandas. 2019. *Python Data Analysis Library*. Available at https://pandas.pydata.org/ [Last accessed on 3 June 2019].

Porter, Shanette C., Michelle Rheinschmidt-Same and Jennifer A. Richeson. 2015. "Inferring Identity From Language". In: *Psychological Science* 27.1, pp. 94–102. DOI: 10.1177/0956797615612202.

Pottinger, A Samuel. 2019(a). *datadrivenempathy/who-wrote-this-news-crawler*. Available at https://github.com/datadrivenempathy/who-wrote-this-news-crawler [Last accessed on 3 June 2019].

Pottinger, A Samuel. 2019(b). *datadrivenempathy/who-wrote-this-server*. Available at https://github.com/datadrivenempathy/who-wrote-this-server [Last accessed on 3 June 2019].

Pottinger, A Samuel. 2019(c). *datadrivenempathy/who-wrote-this-training*. Available at https://github.com/datadrivenempathy/who-wrote-this-training [Last accessed on 3 June 2019].

doi:10.20944/preprints201906.0051.v1

Pottinger, A Samuel. 2019(d). *Machine Learning Techniques for Detecting Identifying Linguistic Patterns in the News Media*. https://www.codeocean.com/. Version v2. DOI: https://doi.org/10.24433/CO.5660509.v2.

Pottinger, A Samuel. 2019(e). "Machine Learning Techniques for Detecting Identifying Linguistic Patterns in the News Media". In: *Protocols.io*. DOI: dx.doi.org/10.17504/protocols.io.3j6gkre. Available at https://www.protocols.io/view/machine-learning-techniques-for-detecting-identify-3j6gkre.

Pottinger, A Samuel. 2019(f). *README*. Available at https://app.wandb.ai/sampottinger/who-wrote-this/reports?view=sampottinger%5C%2FREADME [Last accessed on 3 June 2019].

Pottinger, A Samuel. 2019(g). *Who Wrote This App*. Available at https://whowrotethis.com [Last accessed on 3 June 2019].

Pottinger, A Samuel. 2019(h). *Who Wrote This Data*. Available at https://whowrotethis.com/data [Last accessed on 3 June 2019].

Pottinger, A Samuel. 2019(i). *Who Wrote This Template Workbook*. Available at https://docs.google.com/spreadsheets/d/1ofxgeF6W-G7I6M5Cqcah0d9D-nRDEXpcjKWvekCe7aE/edit?usp=sharing.

Reitz, Kenneth. 2018. *HTTP for Humans*. Available at https://2.python-requests.org/en/master/ [Last accessed on 3 June 2019].

Ribeiro, Marco, Sameer Singh and Carlos Guestrin. 2016. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier". In: *Proceedings of the*

2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. DOI: 10.18653/v1/n16-3020.

Richardson, Leonard. 2019. *Beautiful Soup*. Available at https://www.crummy.com/software/BeautifulSoup/ [Last accessed on 3 June 2019].

Ronacher, Armin. 2019. *Flask (A Python Microframework)*. Available at http://flask.pocoo.org/ [Last accessed on 3 June 2019].

Santamicone, Maurizio. 2019. *Is Artificial Intelligence Racist?* Available at https://towardsdatascience.com/https-medium-com-mauriziosantamicone-is-artificial-intelligence-racist-66ea8f67c7de [Last accessed on 3 June 2019].

Slegg, Jennifer. 2015. *Are There Any SEO Benefits With Having an RSS Feed?* Available at http://www.thesempost.com/are-there-any-seo-benefits-with-having-an-rss-feed/ [Last accessed on 3 June 2019].

SQLite. 2019. *SQLite*. Available at https://www.sqlite.org/index.html [Last accessed on 3 June 2019].

Srivastava, Nitish et al. 2014. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *J. Mach. Learn. Res.* 15.1, pp. 1929–1958. ISSN: 1532-4435. Available at http://dl.acm.org/citation.cfm?id=2627435.2670313.

US Copyright Office. 2019. *More Information on Fair Use*. Available at https://www.copyright.gov/fair-use/more-info.html [Last accessed on 3 June 2019].

Vaux, Bert et al. 2003. *Harvard Dialect Survey*. Available at `http://dialect.redlog.net/` [Last accessed on 3 June 2019].